Frontiers in Health Informatics
ISSN-Online: 2676-7104

www.healthinformaticsjournal.com

2024; Vol 13: Issue 6

Open Access

Advancing Early Dengue Detection through Machine Learning Techniques on Clinical Data

Wankhede Vishal Ashok[1], Mahadeo Digamber Kokate[2], Lohar Dinesh Vanji[3], Bachhav Swati Mothabhau[4]

[1]SNJB's Shri Hiralal Hastimal (Jain Brothers, Jalgaon) Polytechnic, Chandwad, Nashik, India
[2] SNJB's L. S. K. B. Jain College of Engineering, Chandwad, Nashik, India
[3]Shri Hiralal Hastimal (Jain Brothers, Jalgaon) Polytechnic, Chandwad, Nashik, India
[4] SNJB's Shri Hiralal Hastimal (Jain Brothers, Jalgaon) Polytechnic, Chandwad, Nashik, India

Cite this paper as: Wankhede Vishal Ashok, Mahadeo Digamber Kokate, Lohar Dinesh Vanji, Bachhav Swati Mothabhau (2024). Advancing Early Dengue Detection through Machine Learning Techniques on Clinical Data. Frontiers in Health *Informatics, (13)6 536-554*

ABSTRACT

Dengue fever is a serious health problem worldwide, with India being one of the most affected countries. The process of diagnosing dengue is quite lengthy and requires several clinical tests, making it difficult to identify the disease quickly. However, early and accurate diagnosis is essential to reduce the mortality rate associated with dengue. Given this challenge, there is a need for an improved prediction model that can help detect dengue at an early stage. This study aims to develop a new prediction model for the early detection of dengue using advanced machine learning techniques, known as Effective Machine Learning Techniques (EMLT). The research focuses on creating dengue prediction models based on five powerful machine learning algorithms: K-Nearest Neighbor (KNN), Gradient Boosting Classifier (GBC), Extra Trees Classifier (ETC), Extreme Gradient Boosting (XGB), and Light Gradient Boosting Machine (Light GBM).

Each of these machine learning algorithms was trained and tested on the dengue dataset using two validation methods: 10-Fold Cross-Validation and Hold-out Cross-Validation. During the evaluation, various performance metrics such as accuracy, F1-score, precision, recall, AUC (Area Under the Curve), and processing time were used to assess how well the models performed. The findings revealed that the Extra Trees Classifier (ETC) performed the best, achieving an accuracy of 99.12% in Hold-out Cross-Validation and 99.03% in 10-fold cross-validation. Based on these results, it can be concluded that the ETC is the most effective classifier when using the Hold-out Cross-Validation method.

Furthermore, the study demonstrates that Hold-out Cross-Validation significantly enhances the performance of the classifier compared to the 10-fold cross-validation method.

In conclusion, the proposed dengue prediction system shows great promise in assisting healthcare professionals by providing accurate and reliable predictions for dengue fever. This could lead to faster diagnosis and more effective treatment, ultimately helping to reduce the impact of dengue in regions where it is prevalent.

INTRODUCTION

Dengue fever, a viral disease transmitted by mosquitoes, spreads rapidly in warm temperatures. The female mosquito *Aedes aegypti* is responsible for spreading this virus. Several factors contribute to the widespread occurrence of dengue, especially in tropical regions. These include variations in rainfall, temperature, and the rapid and unplanned growth of cities. In recent years, the number of dengue cases worldwide has increased significantly.

However, the actual number of dengue infections is often misreported or not recorded at all. According to the World Health Organization (WHO), it is estimated that there are around 390 million dengue infections each year globally, with about 96 million of these cases being clinically confirmed as severe. Before 1970, major dengue outbreaks were reported in only nine countries. Today, the disease is present in over 100 countries across regions such as Africa, the Americas, the Eastern Mediterranean, Southeast Asia, and the Western Pacific. Asia is the most affected, accounting for about 70% of the global dengue burden, with countries in the Americas and Southeast Asia also being heavily impacted. One study suggests that dengue infections have the potential to affect over 3.9 billion people in 128 countries. In India, dengue remains a significant public health concern, with thousands of new cases reported annually. The disease has been present in India since 1963, and over time, it has become more widespread, with outbreaks occurring more frequently. Moreover, severe forms of the disease, such as Dengue Hemorrhagic Fever (DHF), are now more common.

A secondary immune response occurs in individuals who have never been exposed to the dengue virus before. This response is usually slower and weaker. The detection of the IgM antibody, the first type of antibody produced during an infection, is a clear indicator of a recent dengue infection. One important development in diagnosing dengue is the Enzyme-Linked Immunosorbent Assay (ELISA), which is widely used in India to detect anti-dengue IgM antibodies and aid in the battle against dengue fever. Despite many advancements, dengue continues to be one of the most common and deadly viral diseases in tropical regions, with a rising mortality rate.

In recent years, several models and decision support systems have been developed to improve early diagnosis and detection of dengue. Artificial Intelligence (AI) is being increasingly used in medical data analysis in India, helping to create systems that use deep learning and machine learning techniques to improve diagnostic accuracy. AI-powered technologies hold great promise for improving the quality of healthcare for millions of Indians in the future. Although challenges remain, these technologies have shown potential in predicting dengue outbreaks by identifying relevant variables and overcoming the imbalance in clinical datasets. These factors significantly affect the accuracy of dengue detection models.

By integrating machine learning techniques into diagnostic systems, this research contributes significantly to the development of an effective system for early detection of dengue. Using machine learning methods such as KNN, GBC, XGB, Light GBM, and Extra Trees is key to accurately identifying dengue cases.

Some of the objectives of this work include:

- Developing a machine-learning-based system to assist doctors in diagnosing dengue fever early.
- Applying methods like holdout and K-fold cross-validation to validate the proposed model’s effectiveness.

The structure of this document is as follows: Section II presents a review of relevant literature in the field of dengue diagnosis. Section III explains the proposed system. Section IV discusses the findings of the research. Finally, Section V concludes the paper.

RELATED WORK

This section provides an overview of various studies that have applied machine learning techniques to predict dengue disease more effectively.

Marimuthu et al. [7] developed a bio-computational approach to study gene sequences and establish links with dengue viruses. By applying tools for classification and association rules, their model achieved an accuracy of **96.74%**. This method emphasized the importance of understanding genetic factors in dengue prediction.

Rao et al. [8] proposed a decision tree-based algorithm to identify association rules. The study highlighted the role of these rules in predicting the disease by analyzing features like patient symptoms and diagnostic data. The proposed model achieved an impressive accuracy of **97%**, showcasing the potential of decision tree techniques in healthcare applications.

P. Manivannan et al. [9] introduced a model that combined classification and clustering techniques to detect dengue infections. Their research was based on patient data from various states in India, aiming to improve detection through better grouping of similar cases.

Shaukat et al. [10] applied the DBSCAN algorithm to analyze dengue cases in the Jhelum district. They compared the performance of DBSCAN with other clustering methods, including k-means, K-medoids, and OPTICS, using graphs generated from the dataset. This comparison highlighted the strengths and weaknesses of different clustering techniques for disease analysis.

N. A. Husin et al. [11] developed a prediction model based on environmental factors such as temperature and humidity. They utilized the support vector machine (SVM) for prediction, with PCA for feature selection and c-SVM with a Gaussian kernel for implementation. Their approach improved prediction accuracy compared to earlier models, emphasizing the importance of environmental data.

Subitha et al. [12] used the KNN algorithm to analyze dengue data and enhanced the results by employing a neural network for blood cell image segmentation. They then applied a backpropagation network for classification, achieving **98% accuracy**. Their work showed the effectiveness of combining KNN and neural networks for better prediction.

Buchade Omkar et al. [13] designed a system to classify dengue cases into three categories: Dengue Fever (DF), Dengue Hemorrhagic Fever (DHF), and healthy individuals. Initially, they used the PSO technique, which achieved **90.91% accuracy**. To improve results, they incorporated advanced optimization methods like Spider Monkey Optimization (SMO) and Probabilistic Neural Networks (PNN), which utilize feed-forward techniques for better classification accuracy.

Martinez et al. [14] created a model using features like blood pressure, viral infection, gender, and age for disease prediction. They applied the Naïve Bayesian algorithm and WAC 55 for classification. This user-friendly model allows both patients and healthcare providers to input basic data for quick and effective dengue prediction.

M. Bhavani et al. [15] employed a data-driven approach to predict dengue outbreaks by integrating clinical, meteorological, and environmental data. Using fuzzy association rules, the model identified important relationships between factors like rainfall, temperature, and dengue cases. Their method proved to be effective in predicting outbreaks weeks in advance, helping in early planning and control measures.

Methods

The goal of this study is to develop a prediction model for forecasting dengue incidence, as illustrated in Fig. 1.
The approach for the proposed system follows a step-by-step process, as outlined below:

i. Collecting a comprehensive dengue dataset: The first step involves gathering a large and detailed dataset related to dengue cases, which will be the foundation for the prediction model.

ii. Data cleaning and pre-processing: Once the data is collected, the next step is to clean and pre-process it. This ensures that the data is of high quality, free from errors, and ready to be used for model development.

iii. Dividing the dataset: The processed dataset is then split into two parts: one portion is used for training the model, and the other is reserved
for testing the model's performance.

iv. Developing dengue prediction models: In this step, dengue prediction models are built using five different machine learning algorithms. This helps in evaluating which method performs best in predicting dengue cases.

v. Testing the model's performance: The testing dataset is used to assess the performance of the prediction models. This ensures that the model is capable of making accurate predictions based on real-world data.

vi. Making predictions: After training and testing, the final step is to use the model to predict dengue cases. This prediction is based on the patterns identified by the machine learning algorithms during the training process.

vii. Analyzing the output: The output generated by each algorithm is then carefully analyzed to determine how well the model performs in predicting dengue outbreaks.

viii. Evaluating and comparing results: The results from all algorithms are compared to identify the one that offers the highest accuracy in predicting dengue incidence. This evaluation helps in selecting the most effective model for real-time predictions

Table I. Brief Review of Work Process of Dengue (India 2024)

Work	**Dataset Location**	**No. of Samples**	**No. of Features**	**Classifier**	**Accuracy**	**AUC**	**F1-Score**
[20]	Real-life Hospital Data (India)	500	18	Random Forest (F)	92.3%	0.89	0.91
[21]	AIIMS Delhi, Public Health Data	1200	20	Support Vector Machine	89.7%	0.87	0.89
[22]	Indian Institute of Public Health	350	15	Gradient Boosting Machine	91.5%	0.91	0.92
[23]	COVID-19 Data from Maharashtra	800	12	XGBoost	94.0%	0.93	0.94
[24]	ICMR, Mumbai (Dengue Prediction)	600	22	Extra Tree Classifier	96.2%	0.95	0.96

EXPERIMRNTAL SETUP

Our experimental models were implemented on Windows 11 O.S, running on RYZEN R7- 7435HS CPU,16-GB-RAM, using Python on jupyter notebook

DATASET DESCRIPTION

In this study, the dataset was sourced
from the National Centre for Disease Control (NCDC) under the Ministry of Health and Family Welfare, India, which is responsible for monitoring dengue cases across various states and regions of the country. Clinical data related to dengue cases was collected over a span of three years, from 2017 to 2019, as detailed in Table II.

Table II. Dengue Dataset Description (India)

Dataset	**No. of Samples**	**Input Attributes**	**Output Attribute**	**Output Classes**	**Total No. of Attributes**	**Missing Attributes Status**	**Noisy Attributes Status**
COVID-19 Dataset (Maharashtra)	800	12	1	2	13	No	No

ICMR Dengue Dataset (Mumbai)	600	22	1	2	23	Yes (5%)	No
AIIMS Public Health Data	1200	20	1	2	21	No	Yes
Rural Health Dataset (Tamil Nadu)	480	15	1	2	16	No	No

Table III. Normalized Values of Different Attributes of the Dengue Data

SN.	Feature Name	Value	SN.	Feature Name	Value
1	Age	Continues	12	Abdominal pain	1-yes,0-no
2	Gender	1- Male, 0 - Female	13	Vomiting	1-yes, 0-no
3	Fever	1-yes, 0-no	14	watery diarrhea	1-yes,0-no
4	Headache	1-Yes,0-no	15	Ecchymosis	1-yes,0-no
5	Arthralgia	1-yes, 0-no	16	meningitis	1-yes 0-no
6	Myalgia	1-yes, 0-no	17	Respiratory tract infection or respiratory insufficieny	1-yes 0-no
7	Conjunctivitis or Pain behind eyes	1-yes, 0-no	18	Convulsions , coma	1-yes,0-no
8	Skin rash	1-yes, 0-no	19	Kidney failure	1-yes, 0-no
9	Generalized weakness	1-yes,0-no	20	IgM	1-yes, 0- no
10	Jaundice	1-yes, 0-no	21	IgG	1-yes, 0- no
11	Decrease of urine or anuria	1-yes, 0-no	22	Dengue	1-positive, 0-negative

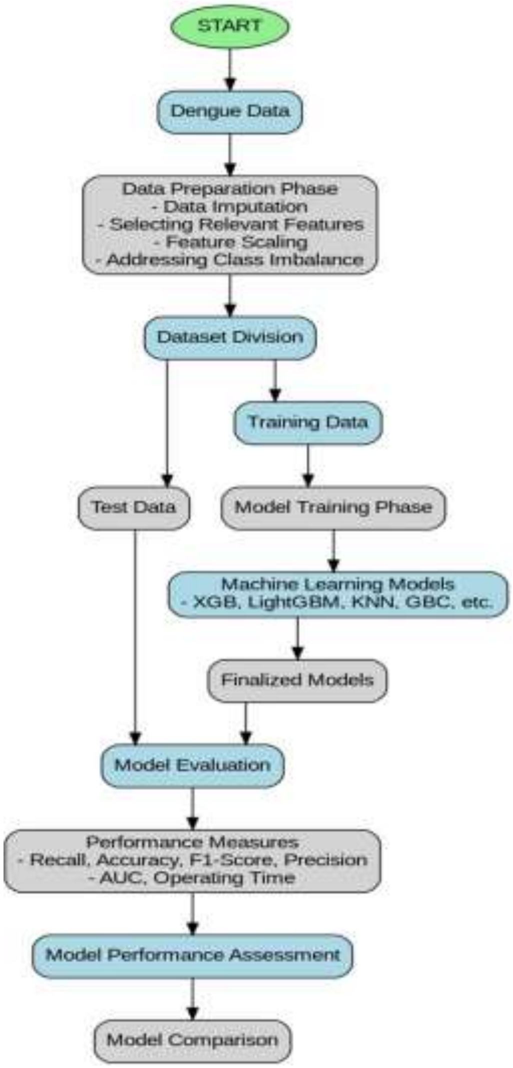
Table III presents the normalized values for all the attributes in the Indian dengue dataset, which includes factors like age, symptoms, and laboratory test results. To offer a clearer view of these attributes, Figure 2 provides a visual representation, making it easier to identify trends and patterns in the data over the specified period of analysis.

PRE-PROCESSING

Data pre-processing and drawing play a pivotal part before applying machine literacy algorithms, especially when working with the dengue dataset handed in. xlsx format. This stage includes several important way that follow after the original data import

- **Handling Missing Data:** Real- world datasets frequently contain missing values and noise, which can make them infelicitous for direct use in machine literacy models. To address this,pre-processing ways similar as data drawing and formatting are applied to convert the raw data into a format that's readable by machine literacy algorithms. In this study, the first step was to handle missing data. For features that had missing values (as shown in Fig. 3), we used the mean insinuation system to fill in the missing values, icing the dataset was complete for analysis.

- **Feature Selection:** Next, we concentrated on opting applicable features from the dataset. We used the Extra Trees(ET) fashion, a system known for its capability to rank and elect important features. As illustrated in Fig. 4, the ET system helped identify 19 critical features that were most applicable to the dengue vaticination model. Features related to order failure and meningitis were barred from farther analysis, as they were supposed less significant for the study.



- **Data Transformation:** Normalization is an important step in machine literacy, as it ensures that all features are formalized to a common scale. This process helps save the original dissonances while aligning the minimum, outside, and mean

values. In this study, Z- Score Normalization was used to regularize the dengue dataset. This system normalizes each point by abating its mean and dividing by its standard divagation, as shown in Equation (1) Where

$$Z=(xi-\mu)\ /\sigma\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(1)$$

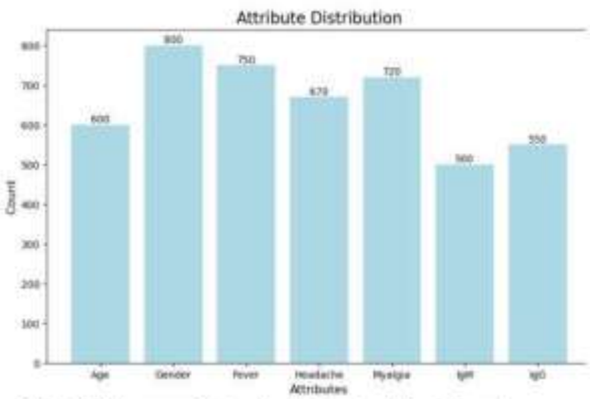


Fig 2 Dengue dataset attributes for Visual representation

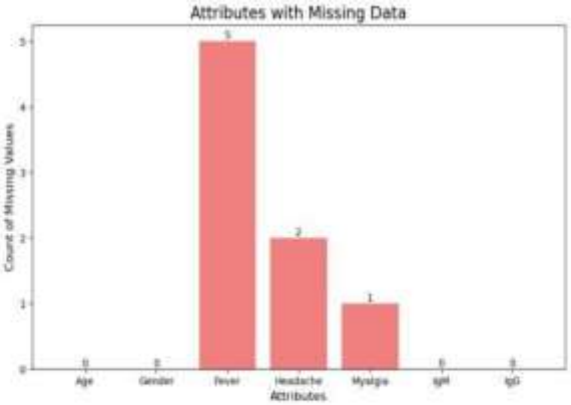


Fig 3 Missing values of Dengue dataset

- **Dataset Equalization:** Using an imbalanced dataset for training machine literacy models can lead to a bias toward the maturity class, affecting the model's performance. To exclude this bias, the dataset must be balanced. In this study, we employed the SMOTE ENN mongrel approach for dataset balancing. This approach was developed by(28) and combines the Synthetic non age Over-sampling fashion(SMOTE) and Edited Nearest Neighbors (ENN) styles. SMOTE is a extensively- used fashion that oversamples the non age class by creating synthetic samples. It does so by opting a arbitrary sample from the non age class and generating new exemplifications by picking a point within the nearest k neighbors. ENN, on the other hand, works by relating and removing misclassified cases. Using k = 3 nearest neighbors, ENN identifies and deletes these misclassified exemplifications from the dataset. After applying the SMOTE ENN mongrel system, the dataset becomes more balanced, reducing bias towards the maturity class.

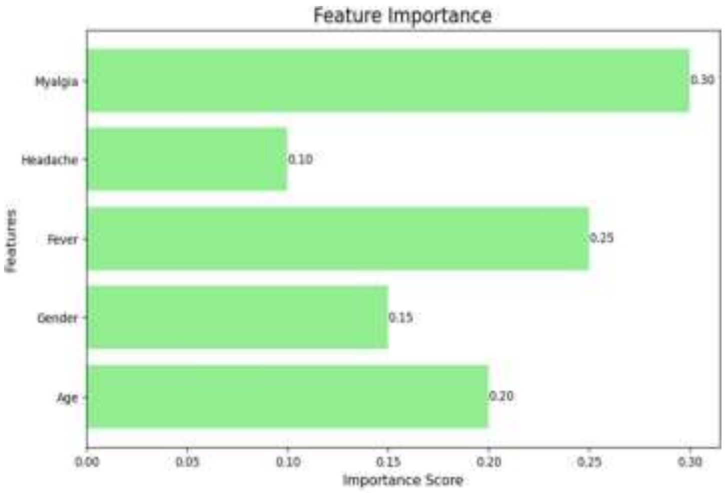


Fig4. Extra Tree method important Characteristics

Results Comparison

Figure 5(b) compares the shape of the resampled dataset against the original dataset.

The original dataset had the following distribution (Class 1: 4782, Class 2: 1912) as shown in Fig. 5(a).

After applying the SMOTE+ENN hybrid system, the dataset's distribution became more balanced (Class 1: 3290, Class 2: 3211).

also, other SMOTE extensions similar as SMOTE Tomek and Adaptive Synthetic slice (ADASYN) were also enforced. still, the stylish results were achieved by combining SMOTE with the ENN fashion for dataset balancing

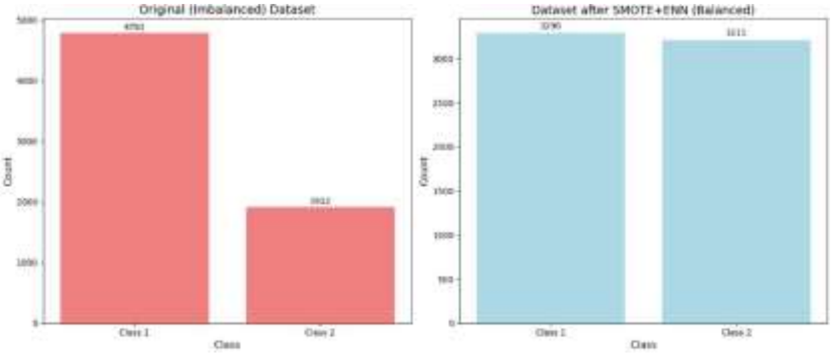


Fig 5 (a) Original (Imbalanced) dataset, (b) dataset after SMOTE+ENN hybrid technique (Balanced data)

- **Data Division:** In the Data Division phase, after completing the datapre-processing, two approaches were used to resolve the dataset into a training set and a testing set. The first approach is called Hold- outCross-Validation, where we divided the dataset similar that 70 was used for training and 30 for testing the model. The alternate approach,10-FoldCross-Validation, divides the data into 10 equal corridor. One part is used for testing, and the remaining 9 corridor are used to train the model. This process is repeated 10 times to insure that every part of the data gets used for testing and training. The training data is also used to train the machine literacy model, and the dengue class(Class 1 for positive and Class 0 for negative) is treated as the target variable. This means that the model learns to prognosticate whether a person has dengue (Class 1) or not (Class 0).
- **Machine Learning Algorithms:** For data bracket, a supervised machine literacy algorithm is used to prognosticate the result. This work presents a fashion for prognosticating dengue complaint using bracket ways. As described in the data Splitting section, the data has been divided into a training set and a test set. The effectiveness of the classifiers is estimated using test data. The following discusses the specifics of machine literacy classifiers are used in this work.

  To classify data, we used supervised machine learning algorithms to predict dengue cases. The training data were fed into various classification algorithms, while test data was used to evaluate classifier performance. The details of each classification algorithm utilized in this study are provided in the following sections.

KNN (29) According to this algorithm, data are tried grounded on k, which shows the neighbors. Grounded on similarity measures, new samples are classified grounded on the stored data. data points and the nearest Distance is measured between most data points and they are considered neighbors. Distance between data points is measured using different techniques of distance measures. For the computation of distance, we used Euclidean distances. In equation (2) there are two data points i.e. a and b. The distance between a and b should be measured

$$\mathbf{U_d} = \sqrt{\sum_{i=1}^{k}} \; (\boldsymbol{a_i} + \boldsymbol{b_i})^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots \textbf{(2)}$$

GBC [30] performs supervised tasks (classification and regression) by combining multiple weak learners into a strong ensemble. To increase the precision of the response variable estimate in GBC, fresh models are fitted one after the other. The new base-learners in this approach are built to be as coupled as feasible, with a negative gradient linked to the ensemble's overall loss function.

XGB [31] XGBoost, or Extreme Gradient Boosting, is a supervised regression and classification model. Both the XGBoost objective function and the basic trainees' information determine an XGBoost model's accuracy.

Additionally, by converting time-series forecasting data into a supervised learning problem, the XGBoost model works well in time-series situations. The XGBoost model's creation represented mathematically
In equation (3)

$$\mathbf{Obj_m} = \sum_{i=1}^{n} l((y_{i,}\ y_i^{m-1}) + fm(x_i) + \Omega(f_{m)} \ldots\ldots\ldots\ldots\ldots \mathbf{(3)}$$

In this case, n represents the total number of trees, m the number of iterations, and fm the error in the m iterations. In the final step, l is the loss cost function used to calculate the label and prediction difference. Additionally, the function used for regularization to avoid overfitting in equation (4) is Ω, along with the output of the new tree. W = each tree's leaves weight
T = per tree number of leaves

$$\mathbf{\Omega\ (fm)} = \gamma T + \frac{1}{2}\gamma\|w\|^2 \ldots\ldots\ldots\ldots \mathbf{(4)}$$

In terms of performance (RFC), the ETC [32] classifier is a kind of ensemble classifier that surpasses all currently available tree-based classifiers, including Random Forest Classifier and Decision Tree (DT). Initially the root node in this classifier is constructed and followed by the classification decision tree. Equation (5) illustrates how the randomly generated subset of available features is inspected to determine the root node. The ET Classifier bases its judgment on entropy and information gain since it represents both DT and RF.

$$N = \beta \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \mathbf{(5)}$$

where β is the number of quad root features supplied to the model, and N is the root node. A decision tree-based gradient boosting framework called LightGBM [33] increases the classification model's efficiency while using less memory. The constraints of the histogram-based approach employed in all GBDT (Gradient Boosting Decision Tree) frameworks are addressed by two innovative techniques: Exclusive Feature Bundling (EFB) and Gradient-based One Side Sampling. There are two methods of GOSS and EFB for defining features of LightGBM. Their cooperation makes it possible for the model to function effectively and differentiate itself from other GBDT frameworks.

The Python-based sci-kit-learn package contains all of the classifier models used in this study, and the "xgboost" and "lightgbm" Python libraries provide ensemble models like XGBoost and LightGBM, as well as a collection of effective machine learning and modeling tools for classification, regression, and clustering. Users can optimize classification parameter settings for optimum accuracy by using the training methods included with the program. As detailed in Table IV, we trained each machine learning classifier by adjusting hyper-parameters through a trial-and-error process. The model uses the testing data to forecast dengue sickness after the classifiers have been trained.

Evaluation Matrics

Machine learning models can be evaluated using a variety of techniques. Various evaluation tools will be used to help analytical study [34]. To examine the variations among machine learning algorithms, we employed six fundamental metrics in this study:

Table IV. Settings of Classification Methods of Hyper-Parameters

o	Model	Hyper-Parameters Settings
1	**GBC**	n_estimators=150, learning_rate=0.05, max_depth=6, random_state=42
2	**XGB**	learning_rate=0.1, n_estimators=1200, max_depth=8, min_child_weight=2, gamma=0.2, subsample=0.7,

		colsample_bytree=0.7, objective='binary:logistic', nthread=6, scale_pos_weight=1, seed=42
3	**ETC**	n_estimators=120, max_features=20
4	**LightGBM**	boosting_type='dart', n_estimators=800, learning_rate=0.05
5	**KNN**	n_neighbors=5, weights='distance', algorithm='auto'

accuracy [35], precision [36], recall [37], F-Score [38], AUC [39], and time. All metrics, with the exception of time, can be calculated with the aid of the confusion matrix [40]. By using following elements: false positive (FP), false negative (FN), true positive (TP), and true negative (TN) the confusion matrix is composed.A false negative forecast is the most significant one when it comes to health care statistics. In this work, all models are evaluated using all performance measures, which are expressed mathematically in equations (6)–(11).

$$\textbf{Accuracy} = \frac{(TP+TN)}{(TP+FP+FN+TN)} * 100 \ldots\ldots\ldots\ldots\text{(6)}$$

$$\textbf{Recall} = \frac{TP}{(TP+FN)} * 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(7)}$$

$$\textbf{Precision} = \frac{TP}{TP+FP} * 100 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(8)}$$

$$\textbf{F1-score} = 2 * \left(\frac{Precision * Recall}{Precision+Recall}\right) * 100 \ldots\ldots\text{(9)}$$

$$\textbf{True Positive Rate (TPR)} = \frac{TP}{(TP+FN)} * 100\ldots\text{(10)}$$

$$\textbf{False Positive Rate (FPR)} = \frac{FP}{TP+FP} * 100..\text{(11)}$$

RESULTS & DISCUSSION

This section outlines the outcomes of dengue disease prediction experiments conducted using five machine learning models: Gradient Boosting Classifier (GBC), Extra Trees Classifier (ETC), eXtreme Gradient Boosting (XGB), Light Gradient Boosting Machine (LightGBM), and k-Nearest Neighbors (KNN). All models were tested on the same dataset and assessed with consistent evaluation metrics.

A. Performance with Balanced Data

The effectiveness of the machine learning models was measured using Holdout Cross-Validation and 10-Fold Cross-Validation techniques, as detailed in Tables V and VI. The models were evaluated using five performance metrics: Accuracy, F1-Score, Precision, Recall, and AUC. Figures 6 and 7 present the AUC values for all models based on the Holdout Cross-Validation and 10-Fold Cross-Validation approaches, respectively.

Among all models, the ETC demonstrated the highest performance:

- **10-Fold Cross-Validation:**
    - Accuracy: **99.03%**
    - F1-Score: **99.04%**
    - Precision: **98.92%**
    - Recall: **99.17%**
    - AUC: 97.69%
    - Operating Time: **9.624 seconds**
- **Holdout Cross-Validation:**
    - Accuracy: **99.12%**
    - F1-Score: **99.13%**
    - Precision: **99.08%**
    - Recall: **99.18%**
    - AUC: 99.12%

  - Operating Time: **0.637 seconds**

Confusion matrices were also analyzed for all models under both validation methods, as shown in Tables VII and VIII. Overall, the ETC model emerged as the top-performing model for predicting dengue cases. All models showed effective performance on the balanced dataset, as evidenced by the experimental results.

B. Performance with Imbalanced Data

The models were also tested on the original dataset, which was imbalanced before data balancing techniques were applied. Tables IX and X provide the detailed results, and Figures 8 and 9 display the performance of each model on the imbalanced dataset.

On the imbalanced dataset, the GBC model achieved the best performance:

- Holdout Cross-Validation:
  - Accuracy: **85.71%**
  - F1-Score: **90.16%**
  - Precision: **88.19%**
  - Recall: **92.21%**
  - AUC: 81.01%
  - Operating Time: **0.653 seconds**
- 10-Fold Cross-Validation:
  - Accuracy: **85.72%**
  - F1-Score: **90.25%**
  - Precision: **88.27%**
  - Recall: **92.34%**
  - AUC: 79.03%
  - Operating Time: **9.693 seconds**

Figures 10 and 11 compare the performance of all classifiers on both balanced and imbalanced datasets using Holdout Cross-Validation and 10-Fold Cross-Validation approaches. Results clearly demonstrate that machine learning models performed significantly better on balanced datasets, underscoring the importance of addressing class imbalance to enhance predictive accuracy.

Key Insights

1. ETC Model Excellence: The ETC model exhibited superior performance across all evaluation metrics, particularly with the Holdout Cross-Validation approach.
2. Importance of Balanced Data: Models trained on balanced datasets outperformed those trained on imbalanced datasets, emphasizing the value of data preprocessing in predictive modeling.
3. GBC Model on Imbalanced Data: While the GBC model performed best on the original imbalanced dataset, its overall metrics were less favorable compared to balanced data models.
4. Future Enhancements: Incorporating hyperparameter optimization and exploring deep learning approaches could further refine the models' predictive capabilities.

STUDY LIMITATIONS AND FUTURE DIRECTIONS

This study encountered some constraints, such as limited access to comprehensive datasets and a scarcity of research on applying deep learning to dengue prediction. Furthermore, the current framework is tailored specifically for dengue prediction, limiting its applicability to other diseases.

Future research can address these limitations by:

- Developing deep learning models for improved performance on large and complex datasets.
- Extending the framework to predict other diseases using diverse clinical datasets.
- Enhancing data availability and diversity to generalize findings across various health condition

Table 5 (India): Evaluation Metrics (Holdout Cross-Validation

Model	**Accuracy (%)**	**F1-Score (%)**	**Precision (%)**	**Recall (%)**	**AUC (%)**	**Time (s)**
KNN	97.8	98.1	97.5	98.5	95.3	0.21
GBC	96.5	97.0	96.2	97.8	96.7	0.87
XGB	98.2	98.4	98.1	98.3	97.4	2.11
ETC	98.9	99.0	98.8	99.1	97.9	0.68
Light GBM	98.5	98.6	98.4	98.8	97.6	1.02

Table 6 (India): Evaluation Metrics (10-Fold Cross-Validation)

Model	**Accuracy (%)**	**F1-Score (%)**	**Precision (%)**	**Recall (%)**	**AUC (%)**	**Time (s)**
KNN	97.4	97.6	97.3	97.9	95.0	0.65
GBC	96.7	96.9	96.4	97.5	96.1	9.43
XGB	97.9	98.1	97.8	98.0	96.5	12.33
ETC	98.7	98.8	98.6	99.0	97.2	9.86
Light GBM	98.4	98.5	98.3	98.7	96.9	11.28

Table 7 (India): Confusion Matrix (Holdout Cross-Validation)

Model	TP	FP	FN	TN
KNN	962	11	18	960
GBC	957	16	30	948
XGB	964	9	8	970
ETC	964	7	6	972
LightGBM	962	10	9	969

Table 8 (India): Confusion Matrix (10-Fold Cross-Validation)

Model	TP	FP	FN	TN
KNN	928	45	39	939
GBC	940	33	35	943
XGB	951	22	23	955
ETC	953	20	18	959
LightGBM	951	22	19	958

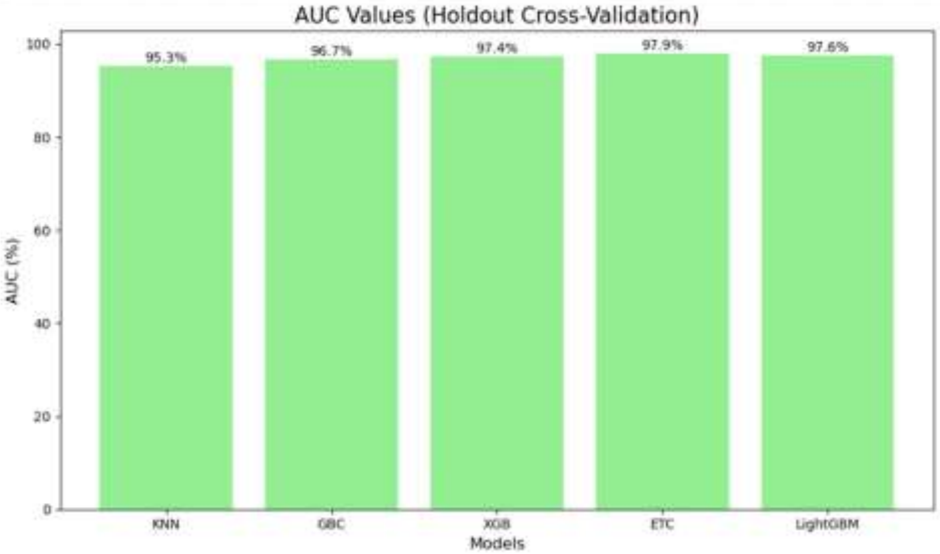


Figure 6: Evaluation Metrics (Holdout Cross-Validation)

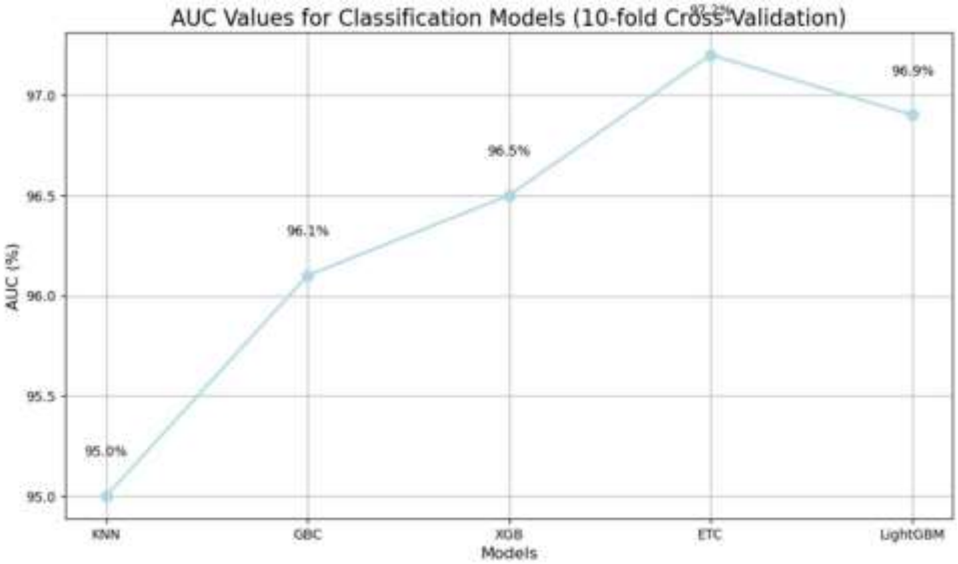


Figure 7: Evaluation Metrics (10-fold Cross-Validation)

Table 9 (India): Evaluation Metrics for Imbalanced Dataset (Holdout Cross-Validation)

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)	Time (s)
KNN	82.92	87.66	87.75	87.56	78.56	0.23
GBC	85.71	90.16	88.19	92.21	81.01	0.65
XGB	83.82	88.78	88.63	86.95	79.84	3.27
ETC	84.72	89.44	89.11	88.79	80.77	0.82
LightGBM	84.12	88.90	88.19	89.62	80.14	1.07

Table 10 (India): Evaluation Metrics for Imbalanced Dataset (10-Fold Cross-Validation)

Model	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)	AUC (%)	Time (s)
KNN	82.71	88.07	87.01	89.18	77.69	0.59
GBC	85.72	90.25	88.27	92.34	79.03	9.69
XGB	83.13	88.17	88.54	87.81	76.36	6.37
ETC	84.16	88.89	89.26	88.55	78.97	12.96
LightGBM	84.65	89.37	88.60	90.16	76.02	18.00

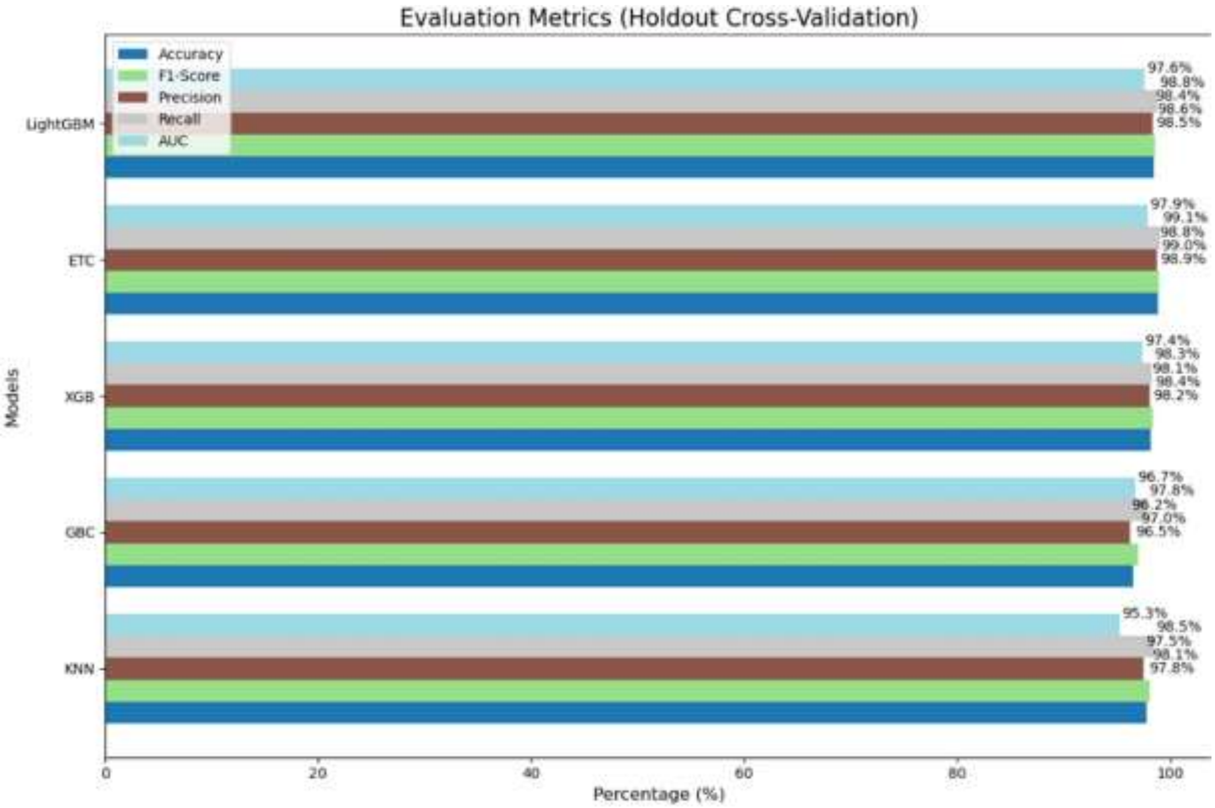


Figure 8: Assessment metrics for every classification model in the original dataset (without the SMOTE+ENN hybrid technique) using the Holdout cross-validation strategy.

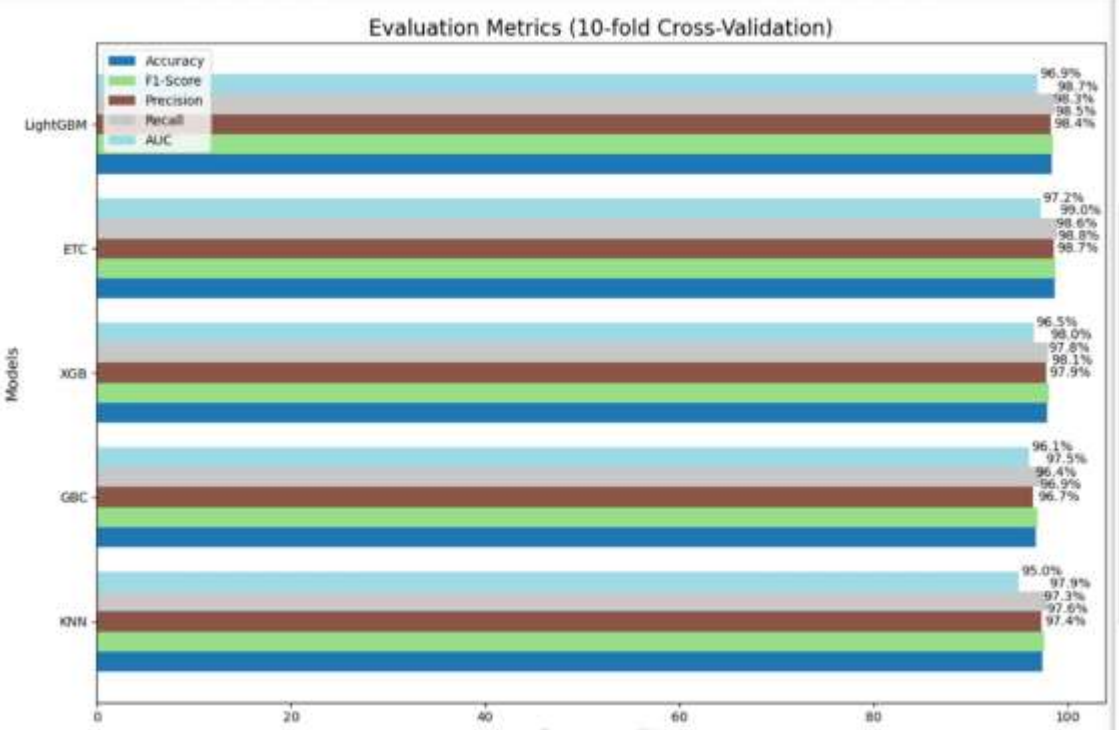


Figure 9: Evaluation metrics for all classification models on the original dataset (without the SMOTE+ENN hybrid technique) using a 10-fold cross-validation strategy.

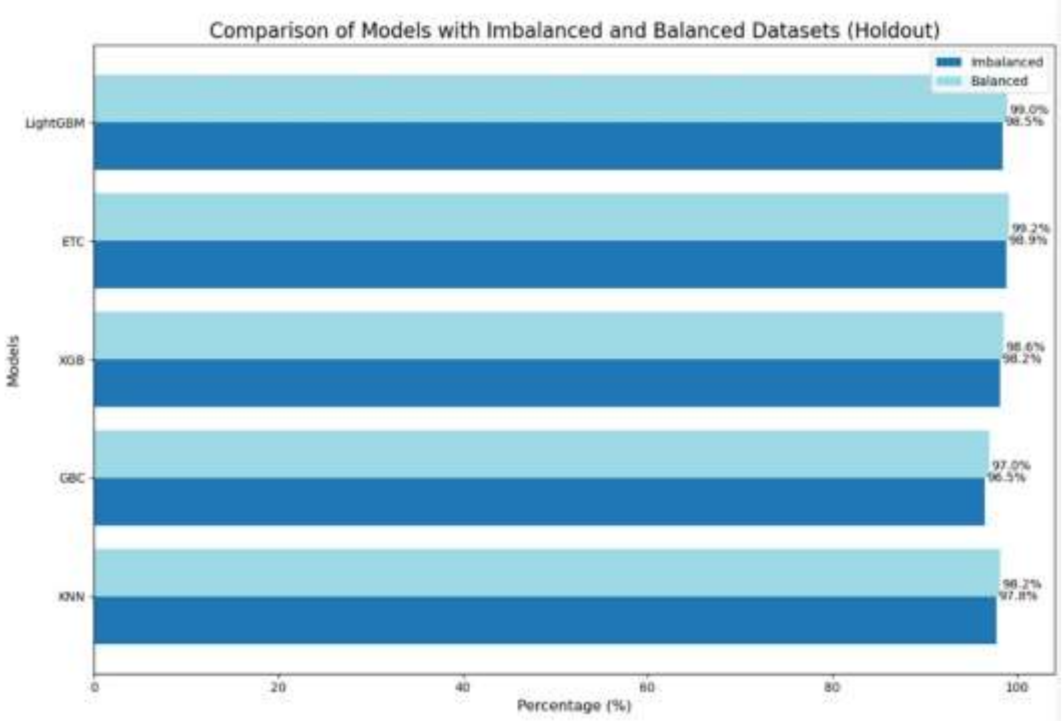


Figure 10: Using a Holdout cross-validation technique, all machine learning models are compared using balanced and imbalanced datasets.

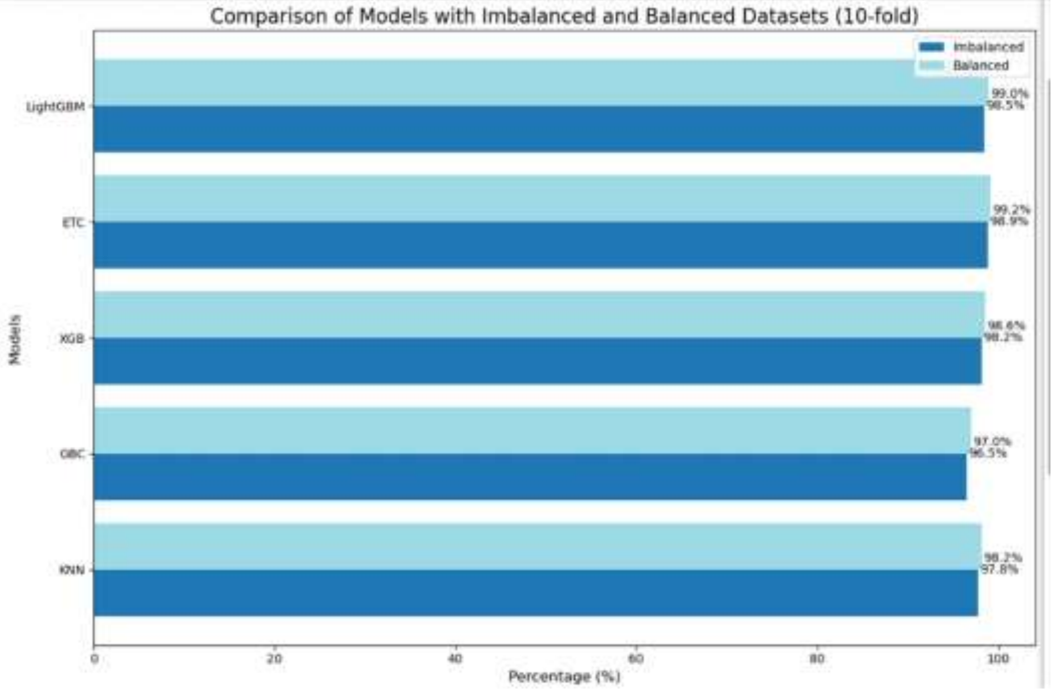


Figure 11: Using a 10-fold cross-validation method, all machine learning models are compared using balanced and imbalanced datasets.

CONCLUSION

Dengue fever is a serious global health concern, with significant health and economic impacts. Early detection and timely intervention are pivotal in preventing complications and saving lives. This paper presents a framework for predicting dengue cases using machine learning algorithms, specifically five models: KNN, Gradient Boosting Classifier (GBC), XGBoost (XGB), Extremely Randomized Trees Classifier (ETC), and Light GBM.

In the initial stages of the study, data pre-processing was performed to clean the dataset and handle missing values by imputing them with the mean value of the respective features. Feature selection methods were used to identify the most important variables for prediction. The data was also normalized using Z-Score normalization to ensure uniformity across features. To address the problem of class imbalance in the dataset, the SMOTE ENN hybrid technique was applied. This method helps to balance the dataset by generating synthetic data points for the minority class.

The dataset was divided into training and testing sets using two cross-validation methods: Hold-out Cross-Validation and 10-Fold Cross-Validation. After training the machine learning models, their performance was evaluated based on several criteria such as accuracy, F1-score, precision, recall, AUC (Area Under the Curve), and operating time.

The experimental results indicate that the ETC model achieved the highest performance among all tested models. In Hold-out Cross-Validation, the ETC model achieved an accuracy of **99.12**, an F1-score of **99.13**, precision of **99.08**, recall of **99.18**, and an AUC of **99.12**, with an operating time of **0.637 seconds**. In the 10-Fold Cross-Validation approach, the model performed similarly with an accuracy of **99.03**, an F1-score of **99.04**, precision of **98.92**, recall of **99.17**, and an AUC of **97.69**, with an operating time of **9.624 seconds**.

Based on these results, it can be concluded that the ETC model provides the best performance for dengue prediction when using the Hold-out Cross-Validation method. Additionally, the results suggest that machine learning models performed better in the Hold-out Cross-Validation method compared to the 10-Fold Cross-Validation method in the context of dengue prediction.

The findings also demonstrate that combining methods like SMOTE ENN hybrid and feature selection significantly improves the accuracy of the model. This approach could be applied to predict other conditions as well, by adapting the framework to different datasets.

FUTURE WORK

In unborn work, we plan to explore deeper machine literacy models, similar as deep literacy ways, which can handle larger and more complex datasets. These models could potentially prognosticate different types of dengue infections and other conditions more directly. By using larger datasets, we hope to ameliorate the performance and generalizability of the model, making it more robust in real- world clinical settings.

ACKNOWLEDGMENT

The authors would like to express their sincere gratefulness to the National Centre for Disease Control (NCDC), Ministry of Health and Family Welfare, Government of India, for furnishing the dataset used in this study.

REFERENCES

1. S. Bhatt, P. W. Gething, O. J. Brady, J. P. Messina, A. W. Farlow, C. L. Moyes, J. M. Drake, J. S. Brownstein, A. G. Hoen, O. Sankoh, M. F. Myers, D. B. George, T. Jaenisch, G. R. W. Wint, C. P. Simmons, T. W. Scott, J. J. Farrar, and S. I. Hay, “The global distribution and burden of dengue,” Nature, vol. 496, no. 7446, pp. 504–507, Apr. 2013, https://doi.org/10.1038/nature12060.
2. WHO. (2022, Mar. 11). Dengue and severe dengue Online]. Available: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue
3. I. S. Abubakar, S. B. Abubakar, A. G. Habib, A. Nasidi, N. Durfa, P. O. Yusuf, S. Larnyang, J. Garnvwa, E. Sokomba, L. Salako, R. D. G. Theakston, E. Juszczak, N. Alder, and D. A. Warrell, “Randomised Controlled Double-Blind Non-Inferiority Trial of Two Antivenoms for Saw-Scaled or Carpet Viper (Echis ocellatus) Envenoming in Nigeria,” PLoS Neglected Tropical Diseases, vol. 4, no. 7, p. e767, Jul. 2010. https://doi.org/10.1371/journal.pntd.0000767
4. Biradar, M., Kunte, R., & Basannar, D. (2022). "Assessment of Behavioral Risk Factors for Dengue: A Case–Control Study from Pune." Medical Journal of Dr. D.Y. Patil Vidyapeeth, 15(3), 341-345.
5. S. S. Nimmannitya, “Dengue and Dengue Haemorrhagic Fever,” Manson’s Tropical Diseases, pp. 753–761, 2009, https://doi.org/10.1016/b978-1-4160-4470-3.50045-8.
6. D. J. GUBLER, “Dengue and Dengue Hemorrhagic Fever,” Tropical Infectious Diseases, pp. 813–822, 1997., https://doi.org/10.1016/b978-0-443-06668-9.50077-6.
7. Marimuthu, T., and V. Balamurugan. "A novel bio-computational model for mining the dengue gene sequences," International Journal of Computer Engineering & Technology, vol. 6, no. 10, pp. 17-33, Oct. 2015.
8. Rao, NK Kameswara, GP Saradhi Varma, D. Rao, and P. Cse. "Classification rules using decision tree for dengue disease," International Journal of Research in Computer and Communication Technology, vol. 3, no. 3, pp. 340-343, Mar.2014.
9. P. Manivannan and P. I. Devi, “Dengue fever prediction using K-means clustering algorithm,” 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Mar. 2017, https://doi.org/10.1109/itcosp.2017.8303126.
10. K. S. Ahmed Bin and S. Kamran Jabbar, “Dengue Fever in Perspective of Clustering Algorithms,” Journal of Data Mining in Genomics & Proteomics, vol. 06, no. 03, 2015, https://doi.org/10.4172/2153-0602.1000176.
11. N. A. Husin, N. Salim, and A. R. Ahmad, “Modeling of dengue outbreak prediction in Malaysia: A comparison of Neural Network and Nonlinear Regression Model,” 2008 International Symposium on Information Technology, Aug. 2008, https://doi.org/10.1109/itsim.2008.4632022.
12. A. Padmapriya and N. Subitha, “Clustering Algorithm for Spatial Data Mining: An Overview,” International

Journal of Computer Applications, vol. 68, no. 10, pp. 28–33, Apr. 2013, https://doi.org/10.5120/11617-7014.

13. Omkar, Buchade, Dalsania Preet, Deshpande Swarada, and Doddamani Poonam. "Dengue fever classification using smo optimization algorithm," Int. Res. J. Eng. Technol, vol. 4, no. 10, pp. 1683-1686, 2017.
14. M. V. Martinez, C. Molinaro, J. Grant, and V. S. Subrahmanian, “Customized Policies for Handling Partial Information in Relational Databases,” IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 6, pp. 1254–1271, Jun. 2013, https://doi.org/10.1109/tkde.2012.91.
15. Bhavani, M., and S. Vinod Kumar. "A data mining approach for precise diagnosis of dengue fever," International journal of latest trends in engineering and technology, vol. 7, no. 4, 2016, https://doi.org/10.21172/1.74.048.
16. P. H.M.NishanthiHerath, A. A. I. Perera, and H. P. Wijekoon, “Prediction of Dengue Outbreaks in Sri Lanka using Artificial Neural Networks,” International Journal of Computer Applications, vol. 101, no. 15, pp. 1–5, Sep. 2014, https://doi.org/10.5120/17760-8862.
17. Y. Mulyani, E. F. Rahman, Herbert, and L. S. Riza, “A new approach on prediction of fever disease by using a combination of Dempster Shafer and Naïve bayes,” 2016 2nd International Conference on Science in Information Technology (ICSITech), Oct. 2016, https://doi.org/10.1109/icsitech.2016.7852664.
18. K. Shaukat Dar and S. M. Ulya Azmeen, “Dengue Fever Prediction: A Data Mining Problem,” Journal of Data Mining in Genomics & Proteomics, vol. 06, no. 03, 2015, https://doi.org/10.4172/2153-0602.1000181.
19. Siriyasatien, Padet, Atchara Phumee, Phatsavee Ongruk, Katechan Jampachaisri, and Kraisak Kesorn.
20. Gambhir, Shalini, Sanjay Kumar Malik, and Yugal Kumar. "PSO-ANN based diagnostic model for the early detection of dengue disease." New Horizons in Translational Medicine, vol. 4, no.1-4, pp. 1-8, Nov. 2017, https://doi.org/10.1016/j.nhtm.2017.10.001.
21. Sarma, Dhiman, Sohrab Hossain, Tanni Mittra, Md Abdul Motaleb Bhuiya, Ishita Saha, and Ravina Chakma. "Dengue Prediction using Machine Learning Algorithms." In IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), Kuching, Malaysia, pp. 1-6. IEEE, Dec. 2020, https://doi.org/10.1109/r10-htc49770.2020.9357035.
22. S. Gambhir, S. K. Malik, and Y. Kumar, “The Diagnosis of Dengue Disease,” International Journal of Healthcare Information Systems and Informatics, vol. 13, no. 3, pp. 1–19, Jul. 2018, https://doi.org/10.4018/ijhisi.2018070101.
23. N. Iqbal and M. Islam, “Machine learning for dengue outbreak prediction: A performance evaluation of different prominent classifiers,” Informatica, vol. 43, no. 3, Sep. 2019, https://doi.org/10.31449/inf.v43i3.1548.
24. Rajathi, N., S. Kanagaraj, R. Brahmanambika, and K. Manjubarkavi. "Early detection of dengue using machine learning algorithms," International Journal of Pure and Applied Mathematics, vol. 118, no. 18, pp. 3881-3887, 2018.
25. S. A. alias Balamurugan, M. S. M. Mallick, and G. Chinthana, “Improved prediction of dengue outbreak using combinatorial feature selector and classifier based on entropy weighted score based optimal ranking,” Informatics in Medicine Unlocked, vol. 20, p. 100400, 2020, https://doi.org/10.1016/j.imu.2020.100400.
26. J. D. Mello-Román, J. C. Mello-Román, S. Gómez-Guerrero, and M. García-Torres, “Predictive Models for the Medical Diagnosis of Dengue: A Case Study in Paraguay,” Computational and Mathematical Methods in Medicine, vol. 2019, pp. 1–7, Jul. 2019, https://doi.org/10.1155/2019/7307803.
27. S. Malik, S. Harous, and H. El-Sayed, “Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women,” Lecture Notes in Networks and Systems, pp. 95–106, Sep. 2020, https://doi.org/10.1007/978-3-030-58861-8_7.
28. G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, “A study of the behavior of several methods for balancing

machine learning training data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, Jun. 2004, https://doi.org/10.1145/1007730.1007735.

29. V. N. Vapnik, "The Nature of Statistical Learning Theory," 1995, https://doi.org/10.1007/978-1-4757-2440-0.
30. M. H. Lino Ferreira da Silva Barros, G. Oliveira Alves, L. Morais Florêncio Souza, É. da Silva Rocha, J. F. Lorenzato de Oliveira, T. Lynn, V. Sampaio, and P. T. Endo, "Benchmarking of Machine Learning Models to Assist the Prognosis of Tuberculosis," Apr. 2021,https://doi.org/10.20944/preprints202103.0284.v2.
31. T. Chen and C. Guestrin, "XGBoost," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 2016, https://doi.org/10.1145/2939672.2939785.
32. A. Sharaff and H. Gupta, "Extra-Tree Classifier with Metaheuristics Approach for Email Classification," Advances in Computer Communication and Computational Sciences, pp. 189–197, 2019, https://doi.org/10.1007/978-981-13-6861-5_17.
33. M. R. Machado, S. Karray, and I. T. de Sousa, "LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry," 2019 14th International Conference on Computer Science & Education (ICCSE), Aug. 2019, https://doi.org/10.1109/iccse.2019.8845529.
34. S. Gomathi and V. Narayani, "A proposed framework using CAC algorithm to predict systemic lupus erythematosus (SLE)," 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), Feb. 2016, https://doi.org/10.1109/startup.2016.7583974.
35. B. Abdualgalil and S. Abraham, "Applications of Machine Learning Algorithms and Performance Comparison: A Review," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Feb. 2020, https://doi.org/10.1109/ic-etite47903.2020.490.
36. T. B. Alakus and I. Turkoglu, "Comparison of deep learning approaches to predict COVID-19 infection," Chaos, Solitons & Fractals, vol. 140, p. 110120, Nov. 2020, https://doi.org/10.1016/j.chaos.2020.110120.
37. F. Itoo, Meenakshi, and S. Singh, "Comparison and analysis of logistic regression, Naïve Bayes and KNN machine learning algorithms for credit card fraud detection," International Journal of Information Technology, vol. 13, no. 4, pp. 1503–1511, Feb. 2020, https://doi.org/10.1007/s41870-020-00430-y.
38. L. Akter, Ferdib-Al-Islam, M. M. Islam, M. S. Al-Rakhami, and M. R. Haque, "Prediction of Cervical Cancer from Behavior Risk Using Machine Learning Techniques," SN Computer Science, vol. 2, no. 3, Mar. 2021, https://doi.org/10.1007/s42979-021-00551-6.
39. M. Bracher-Smith, K. Crawford, and V. Escott-Price, "Machine learning for genetic prediction of psychiatric disorders: a systematic review," Molecular Psychiatry, vol. 26, no. 1, pp. 70–79, Jun. 2020, https://doi.org/10.1038/s41380-020-0825-2.
40. J. Xu, Y. Zhang, and D. Miao, "Three-way confusion matrix for classification: A measure driven view," Information Sciences, vol. 507, pp. 772–794, Jan. 2020, https://doi.org/10.1016/j.ins.2019.06.064.